THE EFFECT OF TOKEN REINFORCEMENT ON STANDARDIZED I.Q. TEST SCORES AS A FUNCTION OF INITIAL I.Q. LEVEL

An abstract of a Thesis by
Therese M. Zylla
November 1980
Drake University
Advisor: Margaret E. Lloyd

The problem. Standardized I.Q. test scores are frequently used as a source of information for making decisions about academic placement (Kolstoe, 1967). Recent research has indicated that individuals differ in motivational level and thus test scores may be reflecting differences in motivation as well as differences in cognitive ability and informational achievements. The present study was concerned with determining whether maximizing motivational level through the use of a token reinforcement program would improve the I.Q. scores of all children or only those children with low I.Q. scores, as previous research has indicated (Clingman & Fowler, 1976).

Procedure. All children were tested on the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) according to standardized instructions. On the basis of the test scores they were placed in either a high I.Q. or a low I.Q. group. Children in each group were randomly assigned to either the experimental or the control condition. Four weeks later, all the children were given a reinforcer effectiveness test. Then the children in the control condition were retested according to standardized instructions. Subjects in the experimental condition were given a token after each correct response. Tokens were exchangable for a variety of activity and tangible items.

Findings. Children in the experimental conditions improved their I.Q. scores significantly over children in the control condition. No interaction was found between condition and I.Q. level. That is, the high experimental group improved over the high control group as much as the low experimental group improved over the low control group.

Conclusions. A token reinforcement program applied contingently for correct responding increased I.Q. scores of preschool children, regardless of initial I.Q. level.

Recommendations. Findings of this study suggest that some children may have motivational deficits and that by maximizing their motivation with a token reinforcement program, their scores may more truly reflect their cognitive ability and informational achievements.

THE EFFECT OF TOKEN REINFORCEMENT ON STANDARDIZED I.Q. TEST SCORES AS A FUNCTION OF INITIAL I.Q. LEVEL

A Thesis

Presented to

The School of Graduate Studies

Drake University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

---

by

Therese M. Zylla

November 1980

THE EFFECT OF TOKEN REINFORCEMENT ON STANDARDIZED I.Q. TEST SCORES AS A FUNCTION OF INITIAL I.Q. LEVEL

by

Therese M. Zylla

Approved by Committee:

Margaret Lloyd
Chairperson

Kipling D. Williams

Raymond A. Hoak

Earle I. Canfield
Dean of the School of Graduate Studies

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The scores produced by standardized intelligence tests are frequently used as a major source of information for making decisions about academic placement within school systems (Kolstoe, 1967) and for personality assessment within clinical practice (Maloney & Ward, 1976) even though the meaning of the scores is unclear. Zigler and Butterfield (1968) suggest that performance on intelligence tests reflects formal cognitive processes, informational achievements and motivational factors. If this is true, then informational achievements and motivational level would have to be kept constant among individuals if I.Q. scores were to measure differences in their cognitive abilities. This problem was acknowledged by Jensen (1980) who pointed out that differences in intelligence are inferred when individuals with equivalent histories perform differently on standardized intelligence tests.

Evidence suggests that testing conditions are not standardized for motivational level and that differences in test scores between individuals may be partially accounted for by differences in motivation (Anastasi, 1954; Eells, Davis, Havighurst, Herrick, & Tyler, 1951; Jensen, 1980; Klugman, 1944; Tiber & Kennedy, 1964; Zigler, 1970). A simple way to rule out motivational discrepancies among

individuals is to maximize motivational levels for all by reinforcing correct answers; for example, some fourth graders' scores on standardized intelligence tests increased when they were given a token for each correct response (Allyon & Kelly, 1972).

Some authors (Clingman & Fowler, 1976; Cohen, 1970; Conner & Weiss, 1974; Jensen, 1980) have suggested that some children are already well motivated in testing situations but that some, usually less economically advantaged children, may not be. If motivation were maximized, then any difference in mean I.Q. between these groups should be reduced.

Headstart children with a mean I.Q. of 82, increased their I.Q. scores by 12 points when candy was given following each correct response (Edlund, 1972); however, the I.Q. scores of white, middle-class children, whose mean I.Q. exceeded 100, did not improve when candy was contingent on correct responses (Clingman & Fowler, 1975). The latter authors suggested that the motivational level for the high I.Q. subjects might already have been optimal and therefore unaffected by reinforcement procedures, whereas the initially low scores of the Headstart children might have been indicative of a motivational deficit that could be eliminated by reinforcement.

This suggestion was tested when children with mean I.Q. scores of 79.3 (low I.Q. group), 100.5 (middle I.Q. group), and 118.5 (high I.Q. group), were given candy

contingent on correct answers to I.Q. test items. Only the children in the lowest third of the distribution increased their scores significantly when reinforced with candy (Clingman & Fowler, 1976). However, the potential effectiveness of candy as a reinforcer for all the children was only established by asking the parents if their children liked candy and by asking the children if they liked candy.

Reinforcers are identified by their effect on behavior, i.e., a positive reinforcer is an event or stimulus which increases the frequency of the response that it follows (Skinner, 1953). Any stimulus or event which does not increase the frequency of the behavior it follows is not a positive reinforcer. Different stimuli function as reinforcers for different individuals. For example, black children's I.Q. scores were shown to increase more when money was the reinforcer than when praise was, while white children's scores increased more when praise was the reinforcer than when money was (Klugman, 1944) and lower-class children learned more quickly when given a candy reward for correct responding while middle-class children performed better when given a neutral feedback condition (light flash for correct responding (Cameron & Storm, 1965).

Generalized conditioned reinforcers, e.g., tokens, are stimuli which acquire their reinforcing properties by being paired with several established reinforcers (Skinner, 1953). Contingent delivery of generalized conditioned

reinforcers may produce similar behavior increases in individuals who have very different preferences in back-up reinforcers. Other advantages of tokens as generalized conditioned reinforcers include the fact that they bridge the delay between the target response and back-up reinforcer delivery, they allow sequences of responses to be reinforced without interruption, and they provide a visible record of improvement (Kazdin & Bootzin, 1972).

It may be difficult to empirically determine the stimuli which function as reinforcers for each of many children. However, it is more probable that a given child's behavior would be changed by the delivery of a token which could be exchanged for a variety of back-up reinforcers than by the delivery of a single item, arbitrarily selected from the back-up reinforcers, e.g., candy. The present study examined the effects of awarding tokens, instead of candy, to children for correct answers on I.Q. test items in order to maximize the motivational level of all the children. Previous research (Clingman & Fowler, 1976) suggests that only low I.Q. children have motivational deficits. This study attempted to determine whether, given a more adequate reinforcement system, high I.Q. as well as low I.Q. children would appear to have motivational deficits and would improve their I.Q. scores.

CHAPTER II

METHOD

Subjects

Seventy children, 4, 5, and 6 years of age, who attended three public day care centers, were tested on the WPPSI. Thirty-two of the children were selected to serve as subjects on the basis of their scores.

Test

Form 1 consisted of the odd numbered items of the WPPSI; form 2 consisted of the even numbered items. Mean split-half reliability coefficients of the verbal, performance and full scale WPPSI I.Q. scores are .94, .93, and .96 respectively (Wechsler, 1967).

Five verbal subtests were selected from the six verbal subtests available, i.e., information, vocabulary, arithmetic, similarities, comprehension and sentences. Split-half reliabilities of the verbal subtests range from .75 to .88 (Wechsler, 1967). The verbal subtest with the lowest split-half reliability for each age group was left out for children in that age group. The comprehension test was omitted for four year olds; arithmetic for 4½ year olds; information for 5 and 5½ year olds; sentences for 6 year olds; and information for 6½ year olds.

Only four performance subtests; picture completion, mazes, geometric design, and block design; were used. The

animal house subtest of the performance scale was not administered since split-half reliability is not considered appropriate for estimating the reliability of speeded tests. Split-half reliabilities of the remaining performance subtests range from .76 to .91 (Wechsler, 1967). The performance I.Q. was prorated according to directions in the manual. The number of consecutive errors a child could make before a particular subtest was discontinued was half the number designated in the test instructions (if the designated number of consecutive errors was odd, the number was rounded upward and then halved, e.g., a test which was normally discontinued after five consecutive errors was discontinued after three consecutive errors).

Test Administration and Scoring

Each child took both forms of the test. Form 2 was administered approximately four weeks from form 1. Both tests were given at the same time of day. The tests were administered by psychology graduate students who had previously taken a testing course. Each graduate student tested an equal number of children in each experimental condition except for adjustments made to accommodate two students who were able to administer form 1 but not form 2. The experimenter calculated subtest raw scores. Raw scores were doubled before being converted into scaled scores and then into I.Q. scores.

Reinforcers and Reinforcer Effectiveness

Prior to the administration of form 2, all the children were asked to draw some circles on a sheet of paper. After thirty seconds each child was given a second piece of paper and told that this time he/she would receive a poker chip for each circle drawn within 30 seconds. The poker chips were exchanged for a variety of back-up reinforcers. Children selected their reinforcers from a prize card, a folder displaying pictures or actual samples of potential reinforcers, i.e., pennies, various flavors of sugarless gum, smiley face and cartoon character stickers, popcorn, raisins, decorative balloons, listening to a story, and playing outside with a grown-up. The number of poker chips each item cost was shown by the number of poker chip sized circles drawn beside it. The back-up reinforcers and their poker chip values are listed in Table 1. The reinforcers themselves (with the exception of playing outside and listening to a story) were stored in the Magic Box, a shoe box wrapped in colorful comic strip paper and tied shut with a brightly colored shoe string. Individual performances on the circle test are shown in Table 2. Some children could earn tokens for correct answers on form 2 of the WPPSI.

Design and Procedure

The odd numbered items of the WPPSI were administered to all seventy children. Children scoring from I.Q. 93

Table 1

Token Values of Back-up Reinforcers

Penny	= 1 token
Sticker	= 2 tokens
Sugarless gum	= 5 tokens
Balloon	= 6 tokens
Raisins	= 6 tokens
Popcorn	= 6 tokens
Story	= 15 tokens
Play outside with adult	= 15 tokens

Table 2

The Number of Circles Drawn by Subjects in both Experimental Conditions during Phase 1 (no Reinforcement for Drawing Circles) and Phase 2 (Token Reinforcement for Drawing Circles) of the Circle Test

Experimental Subjects		Control Subjects	
Phase 1	Phase 2	Phase 1	Phase 2
11	12	6	5
4	5	2	3
5	7	4	5
4	6	12	13
8	11	2	4
5	8	2	4
2	5	4	6
6	10	6	8
2	6	4	6
11	16	10	12
5	10	7	10
5	10	2	5
3	8	3	7
6	12	6	10
3	9	5	10
10	24	3	9

through 111 on the initial test were eliminated from the study. The remaining thirty-two children were evenly divided into a high scoring group with I.Q. scores ranging from 112-139 ($\bar{X}$=119.8) and a low scoring group with I.Q. scores ranging from 61-92 ($\bar{X}$=83.8).

The sixteen subjects in each group were randomly assigned to either an experimental (reinforcement) or a control (no reinforcement) condition. Four weeks after the initial testing on the odd numbered items of the WPPSI all 32 children were retested on the even numbered items. Before taking the second test the children had an opportunity to learn about the value of the poker chips. Each circle they drew in a thirty second period earned a poker chip which was then redeemed according to the values of the items as described on the prize card. Subsequently, children in the experimental condition received a poker chip for each correct response, i.e., one, two, three or four point response. Token delivery was paired with praise, e.g., "wow," "good job." No tokens were given when an incorrect (0 point) response was made by the child. The poker chips were exchanged for the back-up reinforcers shown on the prize card after the test was completed. The tester recorded the items selected by each child on the back of the test. No reinforcement was available to children in the control condition for correct responses.

Reliability

Reliability was taken on the scoring of items on ten tests, five initial tests and five second tests, selected randomly. Reliability was calculated by dividing the number of item score agreements by the number of agreements plus the number of disagreements, and multiplying by 100 (A/A+D x 100). An agreement was scored if both markers awarded an item the same number of points.

A measure of the accuracy of token delivery was calculated for tests administered in the experimental condition. The number of poker chips earned was divided by the number of items scored correct (or vice versa if the number of poker chips earned was greater than the number of correct responses) and multiplied by 100.

CHAPTER III

RESULTS

Reliability of test scoring ranged from .90 to .98, with a mean of .95. Reliability of token delivery ranged from .93 to 1.00, with a mean of .98.

A 2 x 2 analysis of variance showed that high I.Q. subjects had significantly higher I.Q. scores than low I.Q. subjects on both pre and post measures [$F(1,28) = 144.36$, $p<.01$]. There were no significant differences in initial I.Q. levels of the low experimental and low control groups, and no significant differences in initial I.Q. levels of the high experimental and high control groups.

The scores of each child on the first and second test forms and their difference scores are shown in Table 3. Means and standard deviations for each group are also shown in Table 3. The low I.Q. control group had an initial mean I.Q. of 83.5 and a final mean I.Q. of 83.9, showing a mean improvement of .4 points on the second test. The low I.Q. experimental group had an initial mean I.Q. of 84.2 and a final mean I.Q. of 91.6, showing a mean improvement of 7.4 points on the second test. The low experimental group improved 7 points more than the low control group.

The high I.Q. control group had an initial mean I.Q. of 118.7 and a final mean I.Q. of 115.2, showing a mean decrease of 3.5 points on the second test. The high I.Q.

Table 3

Individual Test Scores, Difference Scores, Means and Standard Deviations on Form 1 and Form 2

Low I.Q./No Reinforcement			High I.Q./No Reinforcement		
First Test	Second Test	Change	First Test	Second Test	Change
89	99	+10	114	107	-7
78	72	- 6	120	119	-1
70	60	-10	119	110	-9
78	81	+ 3	115	113	-2
84	82	- 2	114	110	-4
86	86	0	117	120	+3
92	101	+ 9	124	118	-6
91	90	- 1	127	125	-2
$\bar{X}$ 83.5	83.9		$\bar{X}$ 118.7	115.2	
S.D. 7.1	12.7		S.D. 4.2	5.8	

Low I.Q./Reinforcement			High I.Q./Reinforcement		
First Test	Second Test	Change	First Test	Second Test	Change
91	96	+ 5	128	124	- 4
90	96	+ 6	139	133	- 6
89	93	+ 4	119	118	- 1
91	99	+ 8	112	111	- 1
87	90	+ 3	114	128	+14
84	93	+ 9	122	129	+ 7
61	77	+16	112	125	+13
81	89	+ 8	121	130	+ 9
$\bar{X}$ 84.2	91.6		$\bar{X}$ 120.9	124.8	
S.D. 9.4	6.3		S.D. 8.6	6.7	

experimental group had an initial mean of 120.9 and a final mean I.Q. of 124.8, showing a mean improvement of 3.9 points on the second test. The high experimental group improved 7.4 points more than the high control group. Figure 1 shows that the high I.Q. experimental group improved over the high I.Q. control group as much as the low I.Q. experimental group improved over the low I.Q. control group. No condition/I.Q. interaction was found [$F(1,28) = .076$, $p=.785$], contrary to Clingman and Fowler (1976) who reported an interaction, i.e., high experimental subjects improved less than low experimental subjects.

An analysis of variance showed a statistically significant test/condition interaction [$F(1,28) = 11.85$, $p<.002$]. Subjects in the experimental condition increased their scores on the second test significantly more than subjects in the control condition.

Wilcoxin T tests show that differences between the pre and post test scores of the low I.Q. control group ($T=-13.5$, $p>0.05$) and the high I.Q. experimental group ($T=-9$, $p>0.05$) were not significant. Differences between the pre and post test scores of the low I.Q. experimental group ($T=0$, $p<.01$) and the high I.Q. control group ($T=4$, $p<.05$) were significant.

The number of subjects in each group whose scores either increased, decreased, or showed no change from form 1 to form 2 is illustrated in Table 4. Three of the low I.Q.

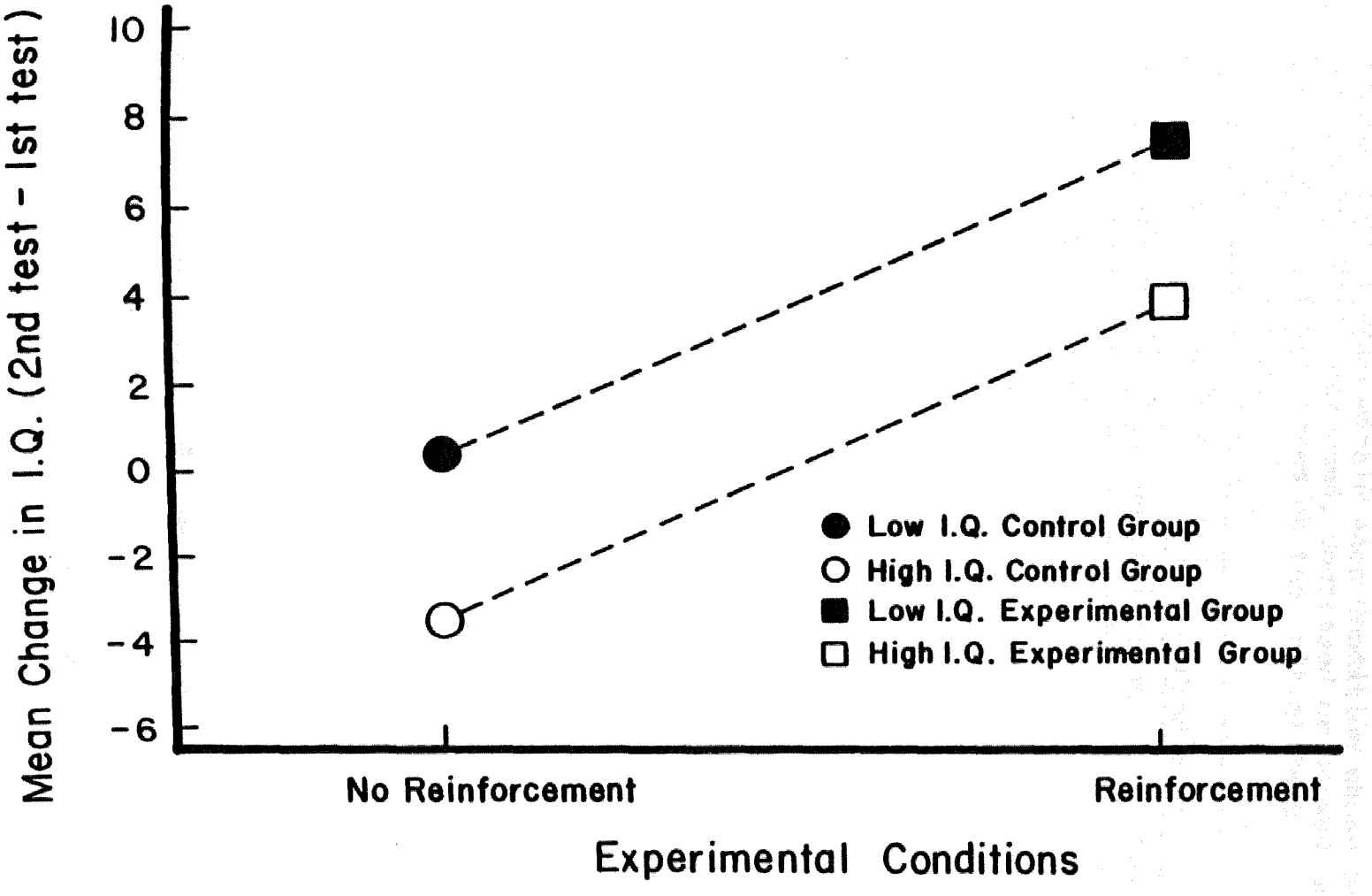

Figure 1. Mean I.Q. change for groups in reinforcement and no reinforcement conditions.

Table 4

Number of Subjects within each Condition whose Scores on the Second Test Increased, Decreased or Showed no Change from Scores on the Initial Test

Initial I.Q. Level	Condition	
	Reinforcement	No Reinforcement
High I.Q.		
Increase	4	1
Decrease	4	7
No Change	0	0
Low I.Q.		
Increase	8	3
Decrease	0	4
No Change	0	1

control group subjects increased their scores, four decreased their scores and one showed no change. All eight of the low I.Q. experimental group subjects increased their scores. One of the high I.Q. control group subjects increased his score and seven decreased their scores. Four of the high I.Q. experimental group subjects increased their scores and four decreased their scores.

Edible reinforcers were selected at least once by 7 children in the low I.Q. group and 8 children in the high I.Q. group. Table 5 shows that 38% of the reinforcers selected by low I.Q. children were edible and 40% of the reinforcers selected by high I.Q. children were edible.

Differences between the number of circles each child drew with and without contingent tokens during the reinforcer effectiveness test were analyzed with a Mann-Whitney U. Children in both reinforcement and non-reinforcement conditions were equally responsive to the reinforcement (U=256).

Table 5

Back-up Reinforcer Selection

High I.Q./Reinforcement Group (n=8)

Back-up Reinforcer	Total Number	Number of Subjects Choosing
Penny	14	6
Gum	13	8
Stickers	13	7
Balloon	12	8
Raisins	9	7
Popcorn	7	6
Play outside	0	0
Story	4	4

Low I.Q/Reinforcement Group (n=8)

Back-up Reinforcer	Total Number	Number of Subjects Choosing
Penny	17	7
Gum	7	5
Stickers	6	3
Balloon	8	5
Raisins	7	6
Popcorn	6	6
Play outside	1	1
Story	1	1

CHAPTER IV

DISCUSSION

Previous research indicated that only children with initially low I.Q. scores ($\bar{X}$=79.3) improved their scores when candy was contingent on correct responding (Clingman & Fowler, 1976). The authors suggested that children with initially high I.Q. scores ($\bar{X}$=118.5) did not increase their scores because they were already working at their maximal motivational level. An alternate explanation of their results is that candy did not function as a reinforcer for all of the children. In the present study children were offered a variety of items and activities as reinforcers for correct test responses. Children with both low ($\bar{X}$=83.8) and high ($\bar{X}$=119.8) scores improved significantly over similar children who were not reinforced for correct responses.

Since both high and low I.Q. children increased their test scores it appears that motivational deficits may exist among children with a considerable range in I.Q.s and may not be characteristic of any one group, i.e., lower I.Q. children. Reducing motivational differences among children would shift the distribution of scores upward, but would not change the mean differences between these groups.

Not all the children in the experimental conditions increased their scores. This may be because not all

children have a motivational deficit in the testing situation or because the range of back-up items was not large enough to ensure that the tokens were reinforcing to all the children. However, it is logical to think that the higher the initial motivational level, the less children's scores should improve under a token reinforcement program. What token reinforcement should do is reduce motivational differences among children so that differences in scores can more safely be attributed to differences in cognitive ability and/or informational achievements.

Although previous research indicates that different groups of children have shown different preferences in reinforcing stimuli, the present study found that there were essentially no differences between back-up reinforcers selected by the high I.Q. and low I.Q. children. However, other studies finding differences in reinforcer preference were not comparing children of different I.Q. levels, but rather, children of different racial groups (Klugman, 1944), or social classes (Cameron & Storm, 1965).

The high control group decreased their I.Q. scores from form 1 to form 2 by a mean of 3.5 points, thus making the high experimental group's mean increase of 3.8 take on greater significance. It is unclear whether this I.Q. score decrease reflects the regression to the mean phenomena or whether it is an artifact of subject selection in this study.

Only when this question is answered and when these procedures have been demonstrated with other age, ethnic, and socioeconomic groups should the standardization of motivational levels as well as testing procedures be considered in the routine administration of I.Q. tests.

REFERENCES

Allyon, T., & Kelly, K. Effects of reinforcement on standardized test performance. Journal of Applied Behavior Analysis, 1972, 5, 477-484.

Anastasi, A. Psychological testing. New York: Macmillan, 1954.

Cameron, A., & Storm, T. Achievement motivation in Canadian, Indian, middle- and working-class children. Psychological Reports, 1965, 16, 459-563.

Clingman, J., & Fowler, R. The effects of contingent and noncontingent reinforcement on the I.Q. scores of children of above average intelligence. Journal of Applied Behavior Analysis, 1975, 8, 90.

Clingman, J., & Fowler, R. The effects of primary reward on the I.Q. performance of grade-school children as a function of initial I.Q. level. Journal of Applied Behavior Analysis, 1976, 9, 19-23.

Cohen, L. The effects of material and non-material reinforcement upon performance of the WISC Block Design subtest by children of different social classes: A follow-up study. Psychology, 1970, 7, 41-47.

Conner, J. J., & Weiss, F. L. A brief discussion of the efficacy of raising standardized test scores by contingent reinforcement. Journal of Applied Behavior Analysis, 1974, 7, 351-352.

Edlund, C. V. The effect on the test behavior of children, as reflected in the I.Q. scores, when reinforced after each correct response. Journal of Applied Behavior Analysis, 1972, 5, 317-319.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. Intelligence and cultural differences. Chicago: University of Chicago Press, 1957.

Jensen, A. R. Bias in mental measurement. New York: The Free Press, 1980.

Kazdin, A. E., & Bootzin, R. R. The token economy: An evaluative review. Journal of Applied Behavior Analysis, 1972, 5, 343-372.

Klugman, S. F. The effect of money incentives versus praise upon the reliability and obtained scores of the Revised Stanford-Binet test. Journal of Genetics and Psychology, 1944, 30, 255-269.

Kolstoe, R. H. Use of test results. Childhood Education, 1967, 44, 165-167.

Maloney, M. P., & Ward, M. P. Psychological assessment, a conceptual report. New York: Oxford University Press, 1976.

Siegel, S. Non-parametric statistics. New York: McGraw-Hill, 1956.

Skinner, B. F. Science and human behavior. New York: The Free Press, 1953.

Tiber, N., & Kennedy, W. A. The effects of incentives on

the intelligence test performance of different social groups. Journal of Consulting Psychology, 1964, 28, 187-189.

Wechsler, D. Manual for the Wechsler Preschool and Primary Scale of Intelligence. New York: Psychological Corporation, 1967.

Zigler, E. The environmental mystique: Training the intellect versus development of the child. Childhood Education, 1970, 46, 402-412.

Zigler, E., & Butterfield, E. C. Motivational aspects of changes in I.Q. test performance of culturally deprived nursery school children. Child Development, 1968, 39, 1-14.